INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System using Deep Recurrent Neural Networks

Cassia Valentini-Botinhao[1], *Xin Wang*[2,3], *Shinji Takaki* [2], *Junichi Yamagishi*[1,2,3]

[1] The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
[2] National Institute of Informatics, Japan
[3] SOKENDAI University, Japan

cvbotinh@inf.ed.ac.uk, {wangxin,takaki,jyamagis}@nii.ac.jp

Abstract

Quality of text-to-speech voices built from noisy recordings is diminished. In order to improve it we propose the use of a recurrent neural network to enhance acoustic parameters prior to training. We trained a deep recurrent neural network using a parallel database of noisy and clean acoustics parameters as input and output of the network. The database consisted of multiple speakers and diverse noise conditions. We investigated using text-derived features as an additional input of the network. We processed a noisy database of two other speakers using this network and used its output to train an HMM acoustic text-to-synthesis model for each voice. Listening experiment results showed that the voice built with enhanced parameters was ranked significantly higher than the ones trained with noisy speech and speech that has been enhanced using a conventional enhancement system. The text-derived features improved results only for the female voice, where it was ranked as highly as a voice trained with clean speech.

Index Terms: speech enhancement, speech synthesis, RNN

1. Introduction

The quality and intelligibility of statistical parametric speech synthesis (SPSS) [1] voices has increased significantly in the last few years. Although adaptation techniques have been shown to improve robustness to recording conditions [2] and despite the wealth of freely available speech data, most studies on SPSS are based on carefully recorded databases. The use of less than ideal speech material, however, is of a great interest for many applications. The creation of personalised voices for instance [3] often relies on recordings that are not of studio quality. Moreover the possibility of using found data, i.e. archived speech material recorded for purpose other than speech synthesis, to increase the amount of training material is quite attractive but restricted by the quality of the recordings. Quality of synthesised speech can be improved by discarding data that is considered to be too distorted by environmental noise, reverberation and microphone quality. When data quantity is small or noise levels are too high discarding seems like a bad strategy. Alternatively speech enhancement can be used to pre-enhance the data. However out of the box speech enhancement methods could distort the signal and speaker characteristics.

Most speech enhancement methods generate either an enhanced version of the magnitude spectrum (or some sort of parametrisation of it) or produce an estimate of the ideal binary mask (IBM) that is then used to enhance the magnitude spectrum [4]. To reconstruct the waveform, phase is derived from the noisy signal or estimated. Recently there has been a strong interest towards statistical-based methods using a deep neural network (DNN) [5, 6, 7, 8]. In [5] a deep feed-forward neural network was used to predict the frequency-domain IBM from noisy spectrum using a cost function in the time domain. A more extensive work on speech enhancement using DNNs is presented in [6] where authors use more than 100 noise types to train a feed-forward network using noise-aware training and global variance [9]. Authors in [7] use text-derived features as an additional input of a feed-forward network that predicts enhanced spectrum parameters and found that distortion is smaller when using text. In most of these studies at least eleven frames (which represent a segment of at least 220 ms) are used as input to the network in order to inform it of the temporal evolution of the features and the context. Alternatively authors in [10, 8] use a recursive neural network (RNN) for speech enhancement. It is worth noting that few studies in the area use wider band speech signals (sampling rate higher than 16 kHz) and most studies only evaluate their systems using objective measures.

There have not been many studies on using speech enhancement for text-to-speech. [11] investigated how additive noise affects feature extraction and the quality of synthetic voices trained using different types of hidden Markov model (HMM) adaptation techniques. Authors found that the excitation parameters are less prone to degradation by noise than cepstral coefficients. They found a significant preference for voices built using clean data for adaptation over voices built with noisy and speech that has been enhanced using a speech enhancement method. More interestingly they found that when the noise is continuous in a database it is better to use a small set of clean data for adaptation than to use a large set of noisy data.

In this paper we propose the use of a recurrent neural network to suppress a range of different additive noises present in a database used to train a text-to-speech system. As proposed in [7] we also investigate the use of text-derived features. Contrary to most speech enhancement methods we directly enhance the acoustic parameters that are used for training the TTS acoustic model, avoiding the unnecessary and error prone reconstruction stage often involved when performing speech enhancement.

This paper is organised as follows: in Section 2 we present a brief summary of RNNs, followed by the proposed speech enhancement system in Section 3 and the experiments in Section 4. Discussions and conclusions follow.

2. Deep recurrent neural networks

RNNs are networks that possess at least one feed-back connection, which could potentially allow them to model sequential data. They are however difficult to train due to the vanish-

http://dx.doi.org/10.21437/Interspeech.2016-159

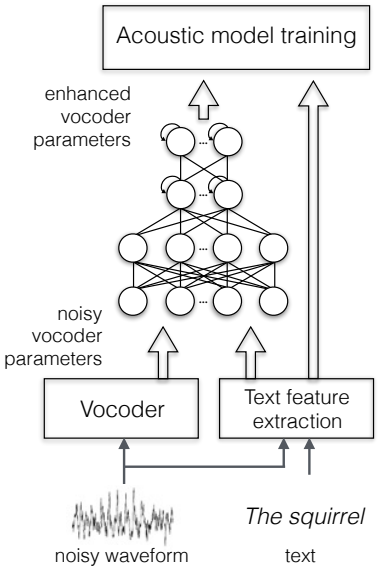

Figure 1: *Proposed speech enhancement.*

ing gradient problem [12]. Long short-term memory networks (LSTM) [13, 14] are recurrent networks composed of units with a particular structure and as such they do not suffer from the vanishing gradient and can therefore be easier to train. An LSTM unit is capable of remembering a value for an arbitrary length of time, controlling how the input affects it, as well as how that value is transmitted to the output and when to forget and remember previous values. LSTMs have been applied in a range of speech problems [15, 16], including regression problems such as text-to-speech [17, 18, 19, 20, 21] and as previously mentioned speech enhancement [10, 8]. LSTMs could be particularly interesting when training with real noisy data, i.e. recordings when speech is produced in noise and therefore changes accordingly.

3. Proposed speech enhancement for TTS

Figure 1 shows the block diagram of how we propose to perform speech enhancement for TTS. Vocoder parameters that describe both the source and the filter are extracted from a time frame of the noisy waveform in the same manner as usually done for TTS training. These parameters are then feed to a neural network together with text-derived features that describe the linguistic context of that particular acoustic frame. The network outputs an enhanced set of acoustic parameters that is then used in conjunction with text-derived features to train a text-to-speech acoustic model. Integrating the speech enhancement as a pre-processing stage while directly enhancing the parameters that are going to be used for training the TTS model avoids unnecessary distortions caused by reconstruction of the waveform. The structure could also be seen as a pre-filter (as opposed to a postfilter that acts at the vocoder level at generation time [22]) and in that sense could potentially be used to minimise synthesis errors as well as enhancement errors.

4. Experiments

In this section we detail the noisy and clean parallel database used for training and testing, the models trained for speech enhancement and text-to-speech, and the design and results of a listening test.

4.1. Database

We selected from the Voice Bank corpus [23] 28 speakers - 14 male and 14 female of the same accent region (England) and another 56 speakers - 28 male and 28 female - of different accent regions (Scotland and United States). There are around 400 sentences available from each speaker. All data is sampled at 48 kHz and orthographic transcription is also available.

To create the noisy database used for training we used ten different types of noise: two artificially generated (speech-shaped noise and babble) and eight real noise recordings from the Demand database [24]. The speech-shaped noise was created by filtering white noise with a filter whose frequency response matched that of the long term speech level of a male speaker. The babble noise was generated by adding speech from six speakers from the Voice Bank corpus that were not used either for either training or testing. The other eight noises were selected using the first channel of the 48 kHz versions of the noise recordings of the Demand database. The chosen noises were: a domestic noise (inside a kitchen), an office noise (in a meeting room), three public space noises (cafeteria, restaurant, subway station), two transportation noises (car and metro) and a street noise (busy traffic intersection). The signal-to-noise (SNR) values used for training were: 15 dB, 10 dB, 5 dB and 0 dB. We had therefore 40 different noisy conditions (ten noises x four SNRs), which meant that per speaker there were around ten different sentences in each condition. The noise was added to the clean waveforms using the ITU-T P.56 method [25] to calculate active speech levels using the code provided in [4]. The clean waveforms were added to noise after they had been normalised and silence segments longer than 200 ms had been trimmed off from the beginning and end of each sentence.

To create the noisy database used for testing we selected two other speakers from England of the same corpus, a male and a female, and five other noises from the Demand database. The chosen noises were: a domestic noise (living room), an office noise (office space), one transport (bus) and two street noises (open area cafeteria and a public square). We used four slightly higher SNR values: 17.5 dB, 12.5 dB, 7.5 dB and 2.5 dB. This created 20 different noisy conditions (five noises x four SNRs), which meant that per speaker there were around 20 different sentences in each condition. The noise was added following the same procedure described previously. The noisy speech database is permanently available at: http://dx.doi.org/10.7488/ds/1356

From the clean and the noisy speech database we extracted using STRAIGHT [26] 60 Mel cepstral coefficients, 25 band aperiodicity components and using SPTK [27] we extracted fundamental frequency (F_0) and voiced/unvoiced information with the RAPT F_0 extraction method [28]. All these features have been extracted using a sliding window of 5 ms shift. We also extracted, using text aligned at a phone level, a 367 dimensional feature composed of: 327 binary, 37 integer and three continuous values. The continuous values represent the length of the current phone and the forward and backward position of the current frame in the phone. The other values encode phone identity of the current frame and the two previous and following phones, as well as a range of other linguistic information usually used for training TTS systems.

4.2. Speech enhancement methods

We used a neural network with two feed-forward layers of 512 logistic units located closest to the input and two bidirectional LSTM (BLSTM) layers of 256 units closest to the output. This

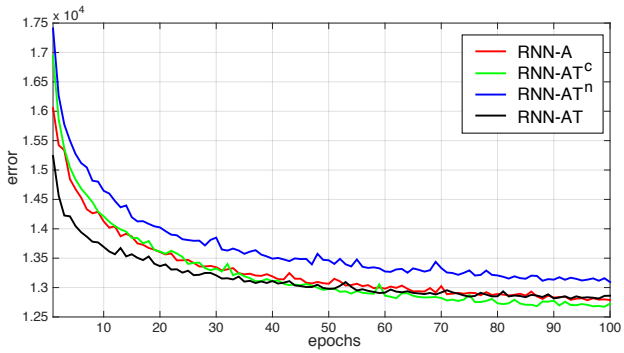

Figure 2: *Validation error during training.*

	MCEP (dB)	BAP (dB)	V/UV (%)	F_0 (Hz)
NOISY	9.86 / 10.68	2.62 / 2.41	9.55 / 7.88	40.27 / 4.38
OMLSA	8.19 / 8.36	3.15 / 2.77	8.73 / 8.28	34.03 / 6.31
RNN-A	4.59 / 5.05	1.86 / 1.72	2.46 / 2.15	24.90 / 8.43
RNN-AT	4.87 / 5.41	1.86 / 1.77	2.61 / 2.25	25.50 / 10.30

Table 1: *Distortion measures calculated from the vocoded parameters used for training the TTS models from a female / male voice.*

configuration was chosen after preliminary experiments with feed-forward only layers and BLSTM only layers showed worse performance with a subset of the data.

The network is trained to map acoustic and text-derived parameters extracted from natural noisy speech to parameters extracted from clean speech. The cost function used is the sum of square errors across all acoustic dimensions. Similar to [10] we set the learning rate to 2.0 e-5 and used the stochastic gradient descent to train the model with randomly initialised weights following a Gaussian distribution with zero mean and 0.1 variance. The momentum was set to zero. We used the CURRENNT tool [29] to train the models using a TESLA K40 GPU board.

The acoustic-only model (RNN-A) was trained with the 56 speaker dataset while the acoustic plus text models were trained with the 28 speaker dataset as the text derived information is dependent on the accent of the speaker. To analyse the effect of adding text-derived features we trained three different models by varying the way we obtain the phone-level alignments. We tested obtaining alignments from the clean as a quality upper bound (RNN-AT^c), noisy as the lower bound (RNN-AT^n) and noisy speech that has been enhanced using the acoustic only (RNN-AT). In the final case we also used the enhanced acoustic parameters as input. For the forced alignment we used an HMM TTS acoustic model that was previously trained with clean data of another speaker. Figure 2 presents the validation error, where we can see that alignment extracted from clean speech obtained lowest errors, followed by enhanced and no text information.

As a conventional speech enhancement method we choose the method described in [30] that uses the optimally-modified log-spectral amplitude speech estimator (OMLSA) and an improved version of the minima controlled recursive averaging noise estimator as proposed in [31]. The code is available from the authors website and has been used as a comparison point for other DNN-based speech enhancement [6].

4.2.1. Objective measures

Table 1 presents distortion measures calculated across the test conditions for each speaker (female/male). We can see that NN-based enhancement decreases mel cepstrum (MCEP), band aperiodicity (BAP) and voiced/unvoiced (V/UV) distortion substantially when compared to the noisy and OMLSA baselines. F_0 distortion measure for the male speaker seems however to increase with all enhancement methods, particularly RNN-based ones. A trend that has also been observed when testing with the same noisy conditions used for training.

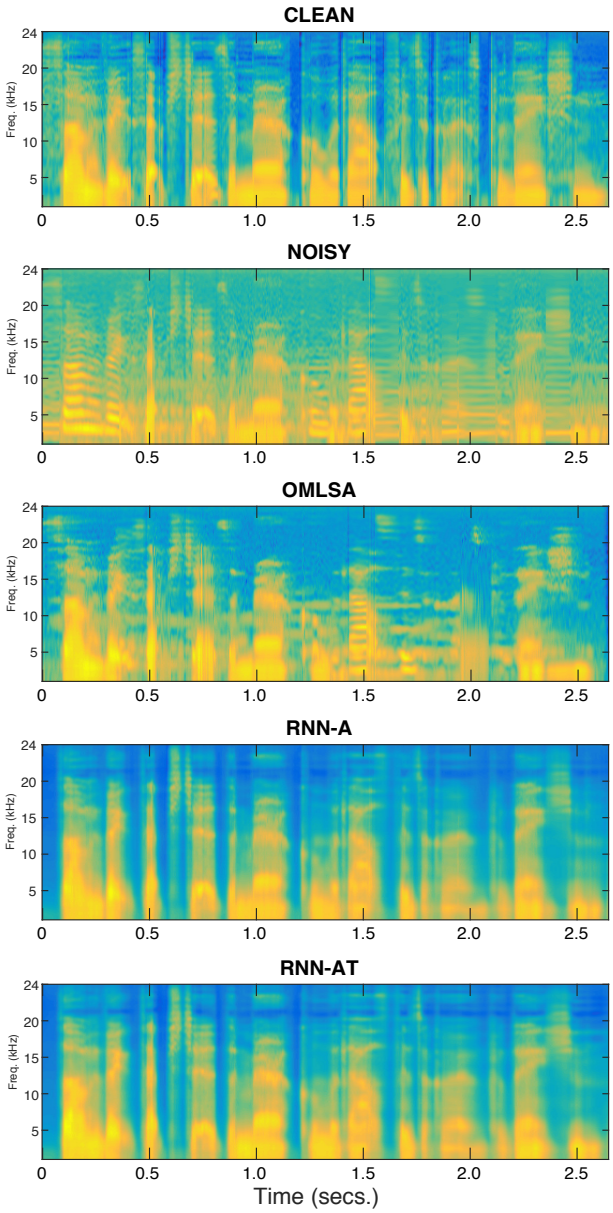

Figure 3: *Spectrum modelled by natural and enhanced Mel cepstral coefficients of speech in noise.*

Figure 3 shows the magnitude spectrum calculated from the Mel cepstral coefficients extracted from natural clean, noisy and OMLSA enhanced speech and generated by the RNN-A and RNN-AT models. The noise in question is the living room noise when the television is on and music is being played. We can see the additive noise has an harmonic structure at some intervals. While the OMLSA enhanced spectrum does still present these erroneous harmonics, see for instance around 1.5 secs, the RNN enhanced spectrum does not. We can see however that the latter appears smoother, something that often occurs when using statistical models for regression.

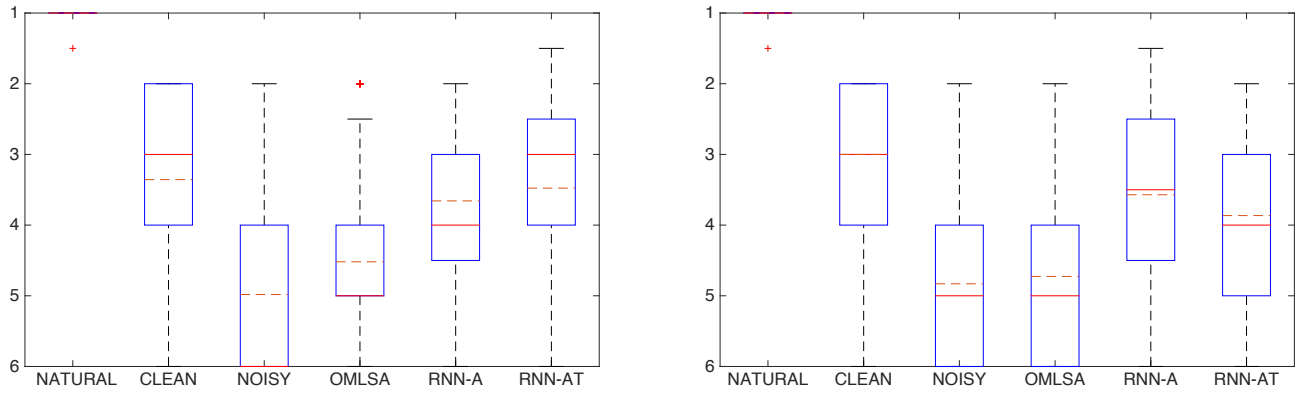

Figure 4: *Rank results of listening experiment with the synthetic female (left) and male (right) voice.*

4.3. Text-to-speech

We built an HMM-based synthetic voice for the female and the male data by adapting the model that was used previously used to obtained the phone-level alignments [32]. Mel cepstral coefficients, band aperiodicities and Mel scale F_0 statics and delta and delta-deltas were used to train the model, forming five streams. To generate from these models we used the maximum likelihood parameter generation algorithm [33] considering global variance [9].

4.4. Listening experiment

We recruited 30 native English speakers to participate on a MUSHRA-style [34] listening test. The test contained 28 screens organised in two blocks: the first 14 with sentences from the male voice and the second half the female voice. The first screen of each block was used to train participants to do the task and familiarise them with the material. In each screen participants were asked to score the quality of a wave sample generated by each system of the same sentence on a scale from 0 to 100. A different sentence is used across different screens. 56 different sentences were used across six listeners. The same sentence was always used for training. Natural speech was also included in the test so that participants would have a reference for good quality as well as checking if participants did go through the material and score it as 100 as instructed. We evaluated five synthetic voices that differ according to the material used for training: clean (CLEAN), noisy (NOISY) and speech data that has been enhanced using OMLSA or the proposed methods RNN-A and RNN-AT.

Figure 4 shows the boxplot of listeners responses in terms of the rank order of systems for the female and the male voice. The order was obtained per screen and per listener according to the scores given to each voice. The solid and dashed lines show medians and means. As a significance test we used the Mann-Whitney U test at a p-value of 0.01 with a Homl Bonferroni correction. As expected, natural speech ranked highest and noise ranked worst. We found that female scores obtained by CLEAN and RNN-AT were not significantly different, as well as RNN-A and RNN-AT scores. The male scores were all significantly different from each other, except OMLSA and NOISY. Each listener uses a different range of the available 0-100 interval and to alleviate the effect that this has we showed the results in terms of rank order. An analysis of raw scores however revealed a similar trend. Scores obtained by RNN-A and CLEAN were however not found to be significantly different, as well as the RNN-AT and RNN-A male scores.

5. Discussion

Objective distortion measures showed that the network trained with acoustic and text features produced higher distortion than the one trained with acoustic features only. Listening test scores for the female voice showed a different trend: the synthetic voice trained with speech enhanced using acoustics and text was scored slightly higher and it was not rated significantly different from the voice trained using clean speech. Text-derived features did not improve quality of the synthetic male voice, but for that particular voice all enhancement methods performed worse. We believe this could be due to F_0 extraction and alignment issues. Although F_0 errors in the noisy female data were greater, the enhancement worked as intended. In the male voice case the errors were so small that the enhancement resulted in extra noise. Since the errors were gender specific we tested with a different network for each, however that did not decrease the F_0 errors of the male voice. Still the fact that the OMLSA, without direct F_0 extraction, did not do any better for the male voice seems to be an indication that particular condition was more challenging regardless. We observed that validation error was lower when using the clean data for alignment, which indicates that results could be further improved with a better alignment. Preliminary tests using an automatic speech recogniser showed less discrepancies between alignment obtained with clean and noisy data.

6. Conclusion

We proposed the use of a recurrent neural network to enhance acoustic features extracted from noisy speech data. We have found that quality of synthesised speech produced with models that have been trained with enhanced data is significantly better and for a particular speaker was not significantly different from the models trained with clean data. This study focused on using HMM-based TTS acoustic models so the question whether similar results would be seen when using DNN-based models remains. This is of particular interest as a more extensive work on DNN adaptation with noisy speech does not exist. Another question that remains open is how these results compare to DNN-based speech enhancement methods that perform waveform reconstruction. Testing with real recordings of noisy data is also of interest.

Acknowledgements This work was partially supported by EPSRC through Programme Grant EP/I031022/1 (NST) and EP/J002526/1 (CAF) and by CREST from the Japan Science and Technology Agency (uDialogue project). The full NST research data collection may be accessed at http://hdl.handle.net/10283/786.

7. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Comm.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] J. Yamagishi, Z. Ling, and S. King, "Robustness of HMM-based Speech Synthesis," in *Proc. Interspeech*, Brisbane, Australia, Sep. 2008, pp. 581–584.

[3] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction," *J. of Acoust. Science and Tech.*, vol. 33, no. 1, pp. 1–5, 2012.

[4] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 1st ed. Boca Raton, FL, USA: CRC Press, Inc., 2007.

[5] Y. Wang and D. Wang, "A deep neural network for time-domain signal reconstruction," in *Proc. ICASSP*, April 2015, pp. 4390–4394.

[6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan 2015.

[7] K. Kinoshita, M. Delcroix, A. Ogawa, and T. Nakatani, "Text-informed speech enhancement with deep neural networks," in *Proc. Interspeech*, Sep. 2015, pp. 1760–1764.

[8] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Roux, J. R. Hershey, and B. Schuller, *Proc. Int. Conf. Latent Variable Analysis and Signal Separation*. Springer International Publishing, 2015, ch. Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR, pp. 91–99.

[9] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

[10] F. Weninger, J. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. GlobalSIP*, Dec 2014, pp. 577–581.

[11] R. Karhila, U. Remes, and M. Kurimo, "Noise in HMM-Based Speech Synthesis Adaptation: Analysis, Evaluation Methods and Experiments," *J. Sel. Topics in Sig. Proc.*, vol. 8, no. 2, pp. 285–295, April 2014.

[12] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *J. Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *J. Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[15] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. ICASSP*, 2013, pp. 6645–6649.

[16] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *CoRR*, vol. abs/1402.1128, 2014.

[17] S.-H. Chen, S.-H. Hwang, and Y.-R. Wang, "An RNN-based prosodic information synthesizer for mandarin text-to-speech," *Proc. ICASSP*, vol. 6, no. 3, pp. 226–239, 1998.

[18] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks." in *Proc. Interspeech*, 2014, pp. 1964–1968.

[19] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks." in *Proc. Interspeech*, 2014, pp. 2268–2272.

[20] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. ICASSP*. IEEE, 2015, pp. 4470–4474.

[21] S. Achanta, T. Godambe, and S. V. Gangashetty, "An investigation of recurrent neural network architectures for statistical parametric speech synthesis," in *Proc. Interspeech*, 2015.

[22] L.-H. Chen, T. Raitio, C. Valentini-Botinhao, Z. Ling, and J. Yamagishi, "A deep generative architecture for postfiltering in statistical parametric speech synthesis," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 23, no. 11, pp. 2003–2014, 2015.

[23] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. Int. Conf. Oriental COCOSDA*, Nov 2013.

[24] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 3591–3591, 2013.

[25] *Objective measurement of active speech level ITU-T recommendation P.56*, ITU Recommendation ITU-T, Geneva, Switzerland, 1993.

[26] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Comm.*, vol. 27, pp. 187–207, 1999.

[27] *Speech signal processing toolkit: SPTK 3.4*, Nagoya Institute of Technology, 2010.

[28] D. Talkin, "A robust algorithm for pitch tracking," *Speech Coding and Synthesis*, pp. 495–518, 1995.

[29] F. Weninger, "Introducing CURRENNT: The Munich Open-Source CUDA RecurREnt Neural Network Toolkit," *J. of Machine Learning Research*, vol. 16, pp. 547–551, 2015.

[30] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403 – 2418, 2001.

[31] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sept 2003.

[32] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 1, pp. 66 –83, 2009.

[33] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, vol. 3. IEEE, 2000, pp. 1315–1318.

[34] *Method for the subjective assessment of intermediate quality level of coding systems*, ITU Recommendation ITU-R BS.1534-1, International Telecommunication Union Radiocommunication Assembly, Geneva, Switzerland, March 2003.